Finding Genes

Finding Genes

1. RNASeq experiments
2. Evolutionary homology
3. Theoretical prediction

RNASeq experiments

1. Collect mRNA
2. Subtract rRNA
3. Construct cDNA
4. Sequence
5. Assemble
6. Map

RNASeq experiments

1. Collect mRNA
2. Subtract rRNA
3. Construct cDNA
4. Sequence
5. Assemble
   - To determine the genes the reads must be assembled.
   - This is commonly done via finding overlap between the reads.
   - Common assembly programs that are exon/intron aware are `Trinity`, `Velvet` and `TopHat`.
6. Map

RNASeq experiments

1. Collect mRNA
2. Subtract rRNA
3. Construct cDNA
4. Sequence
5. Assemble
6. Map
   - To determine where the reads are in a genome the reads must be mapped back to the genome.
   - Common mapping programs are `Cufflinks` and `BWA`.

RNASeq experiments

Reads often show that some regions that are transcribed but have never been annotated as a gene.

Sometimes transcripts are difficult to interpret. They might begin long before the gene. They might end long after the gene. Some transcripts might not correspond to genes at all.

Finally, many genes are not highly expressed.

Homology

Homology is a critical measure for the presence of a gene.

1. Evolutionary conservation
2. Phylogenetic footprinting
3. Phylogenetic co-occurrence
4. Phylogenetic co-evolution

Homology

1 Evolutionary Conservation

Important functional regions are known to evolve slowly and hence to be similar in related organisms.

For some particularly slowly changing genes this relationship can be quite distant.

The similarity and measuring rates of change and finding them to be slow are an indication of a gene or other functional region.

2 Phylogenetic footprinting

Phylogenetic footprinting is a method for the discovery of regulatory elements or another functional pattern in a set of ortholgous regulatory regions from multiple species. It does so by idenitfying the best conserved motifs in those orthologous regions.

Idea of phylogenetic footprinting originated in 1988 by Tagle et al.

Homology

2 Phylogenetic footprinting

Functional sequences evolve at a slower rate than non-functional sequences because of the selective pressure.

Phylogenetic footprinting exploits the mutation rate difference in orthogolous sequence between funcitonal and non-functional sequences.

`FootPrinter` and `BigFoot` are common programs to aid this analysis.

Phylogenetic footprinting is best for finding short elements; regulatory regions, small exons, binding regions.

Homology

3. Phylogenetic co-occurrence

Often bacteria lose/gain one or more genes. If the regions occur together it suggests a functional relationship. Best for functional analysis.

4. Phylogenetic co-evolution

Genes with related functions will evolve together. Again best for functional analysis.

Homology

Other homology uses:

Often it is useful to include EST's, RNAseq data and raw sequences from related organisms in the mapping phase. The evolutionary homology is likely to point to similar genes within the organism of interest.

Prediction

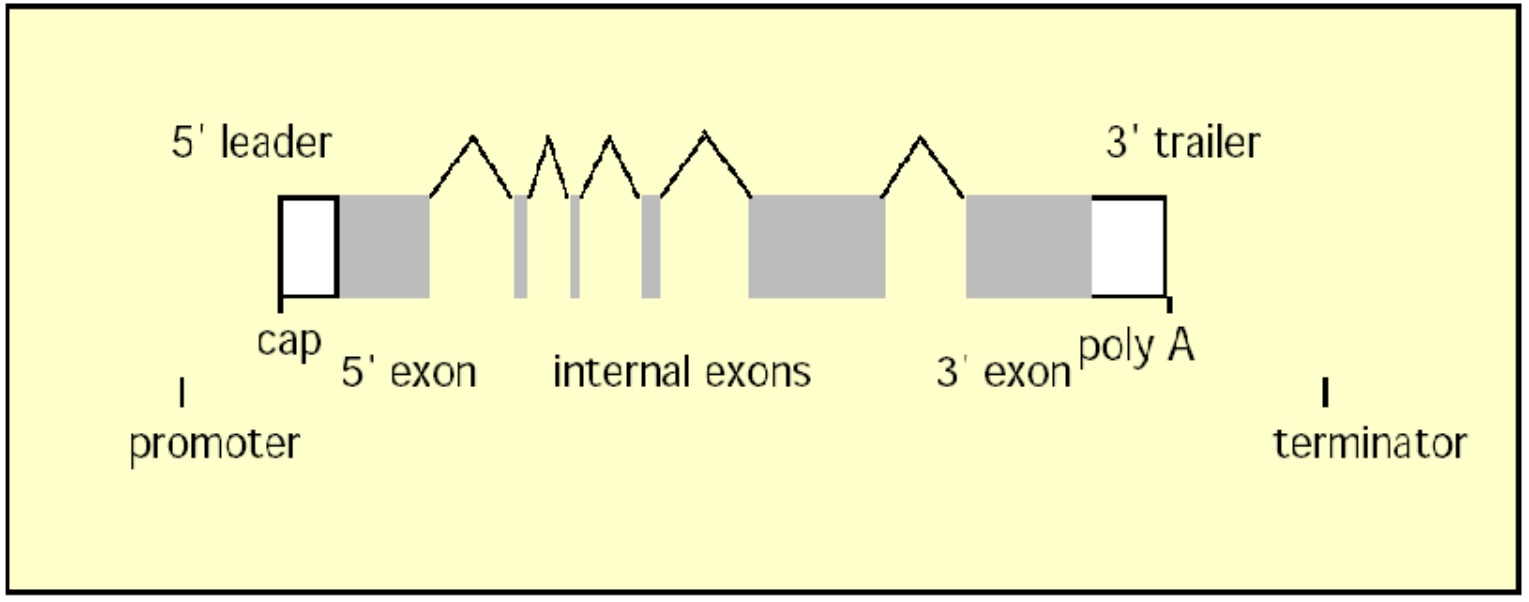
Genes are a "pattern". Given the machinery of the previous chapter we can find these patterns. Some potential patterns are ...

1. Open reading frames
2. codon usage
3. amino acid usage
4. base pair periodicity
5. intron/exon splicing patterns
6. 6-mer frequencies.

Open Reading Frames

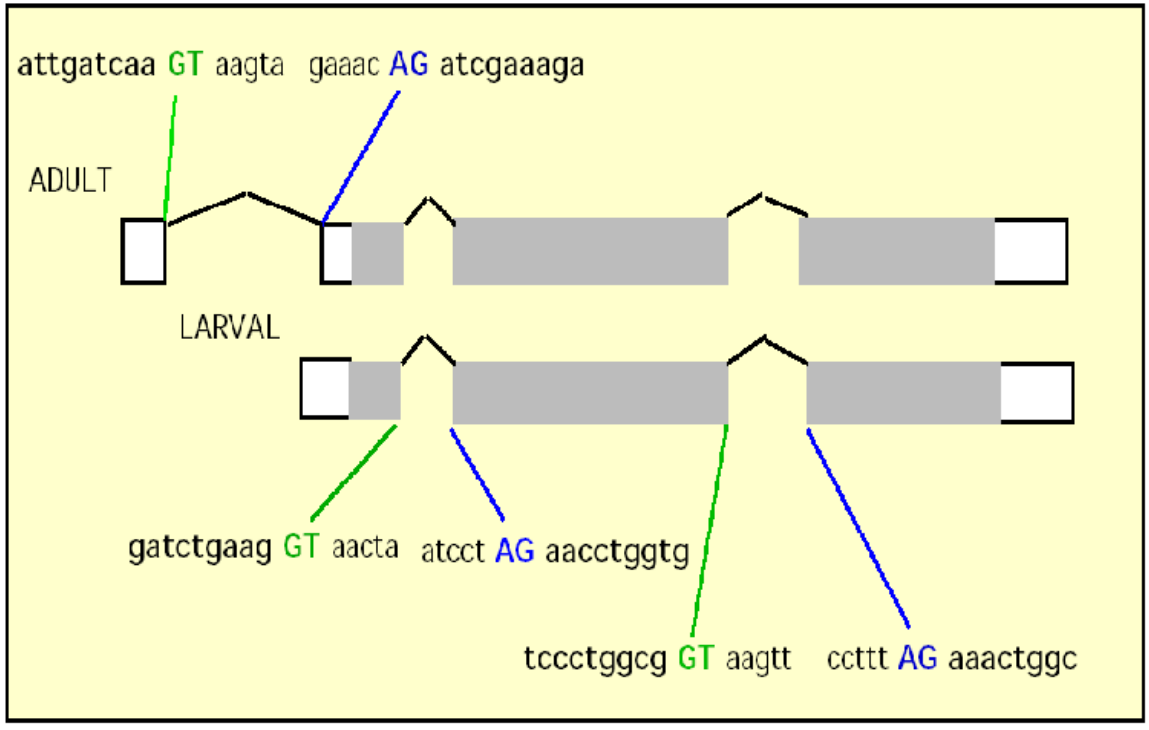
In eukaryotes open reading frames (while necessary) are not that useful for the detection of genes due to the ubiquity of comparatively small exons.

Gene Recognition



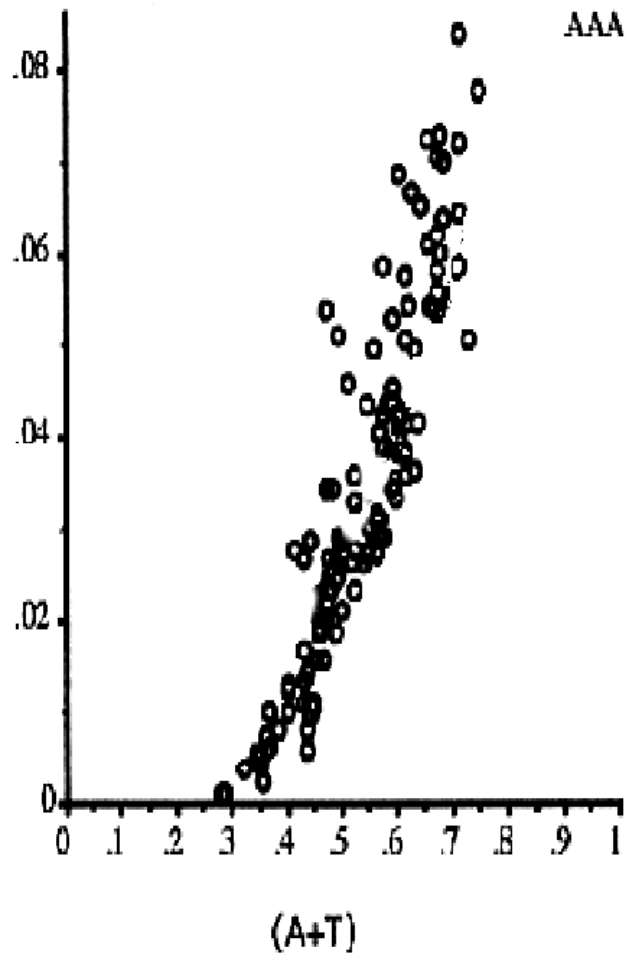
Eukaryotic gene with exon - intron structure, protein-coding is gray.

Splice Sites



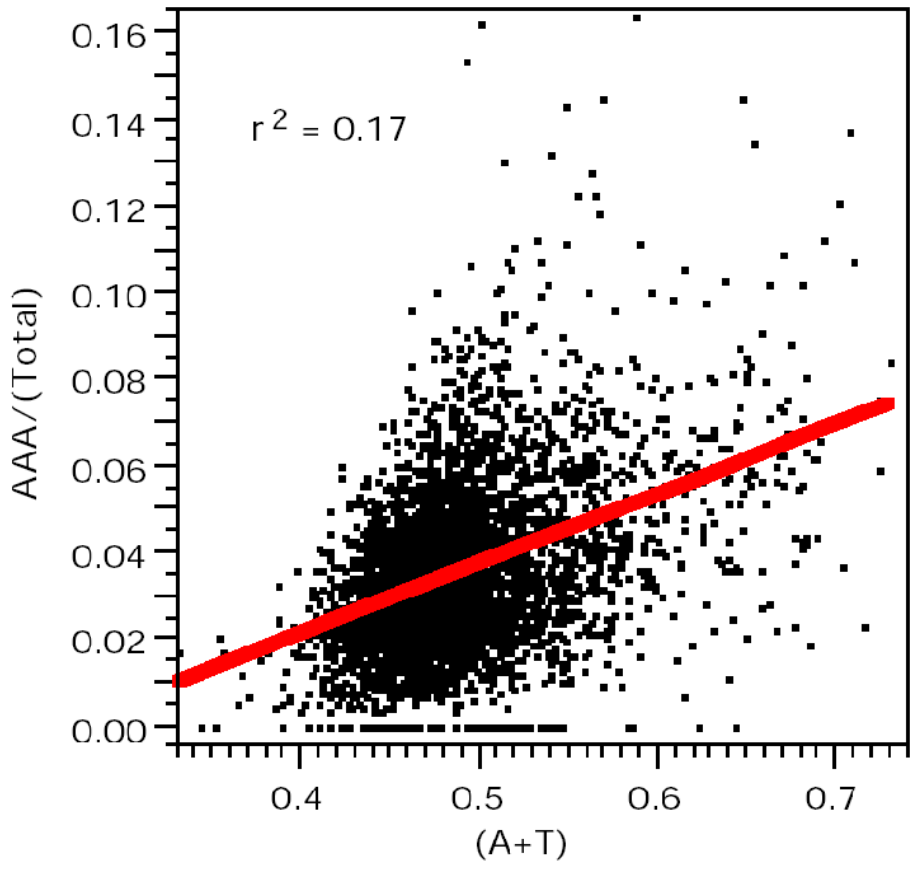
Exon - intron boundaries of the *D. melanogaster* Adh gene.

Codon Usage



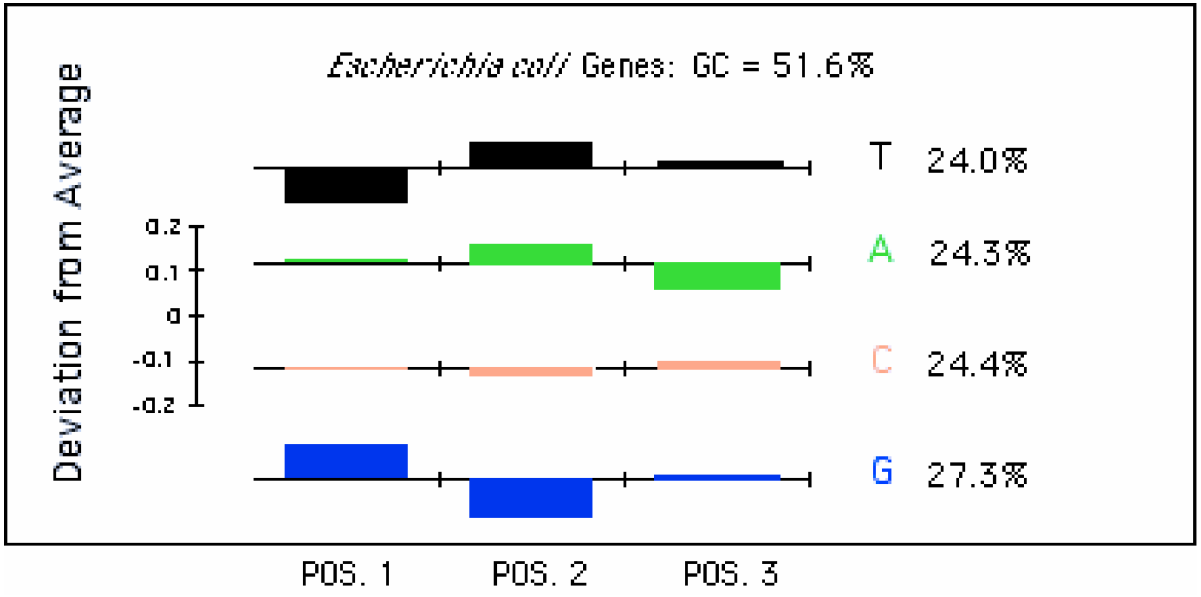
The fraction of all codons that are AAA across genomes with different AT contents.

Codon Usage



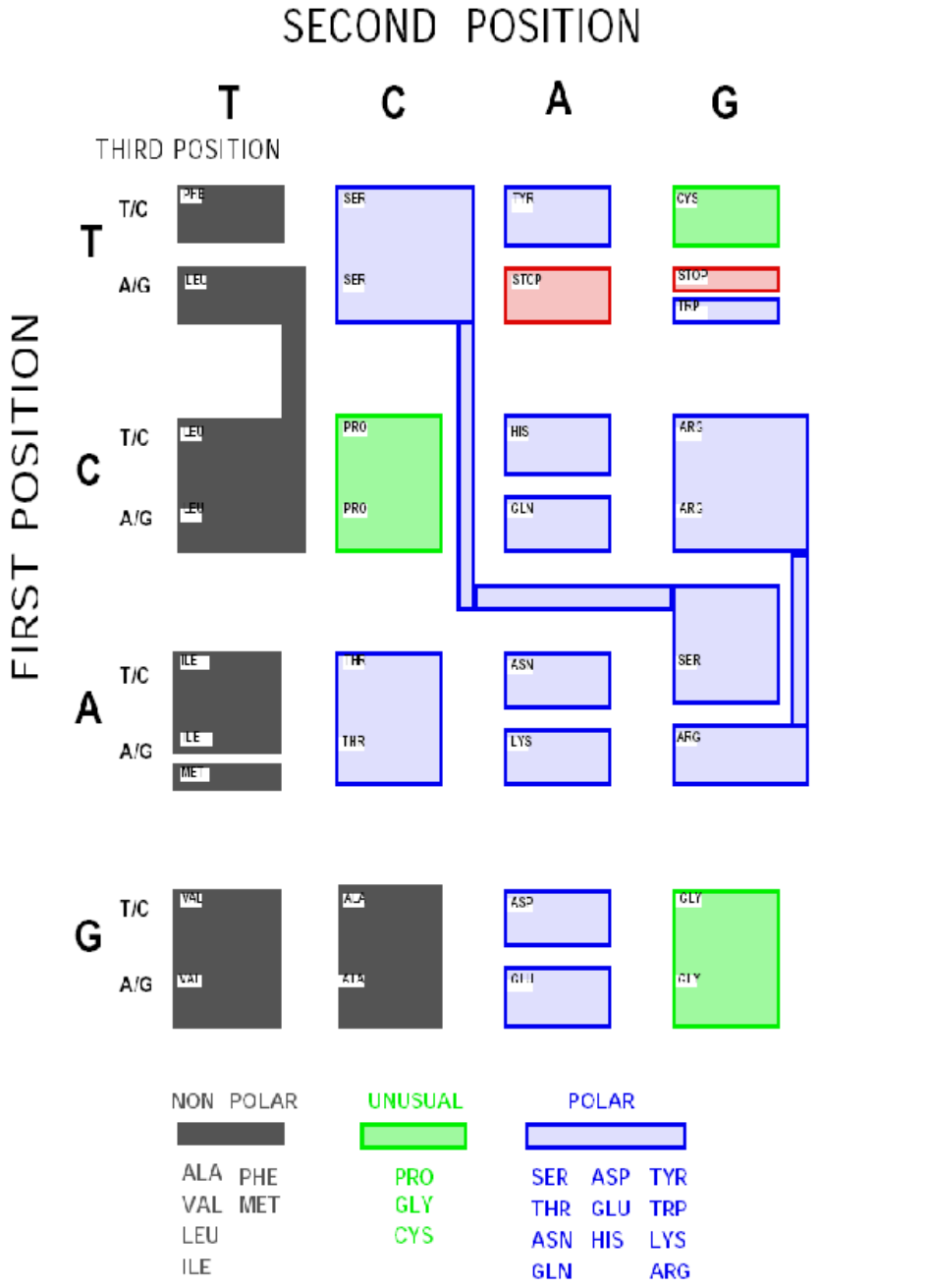
The fraction of codons that are AAA for genes of the *E. coli* genome as a function of AT concentration

Codon Usage

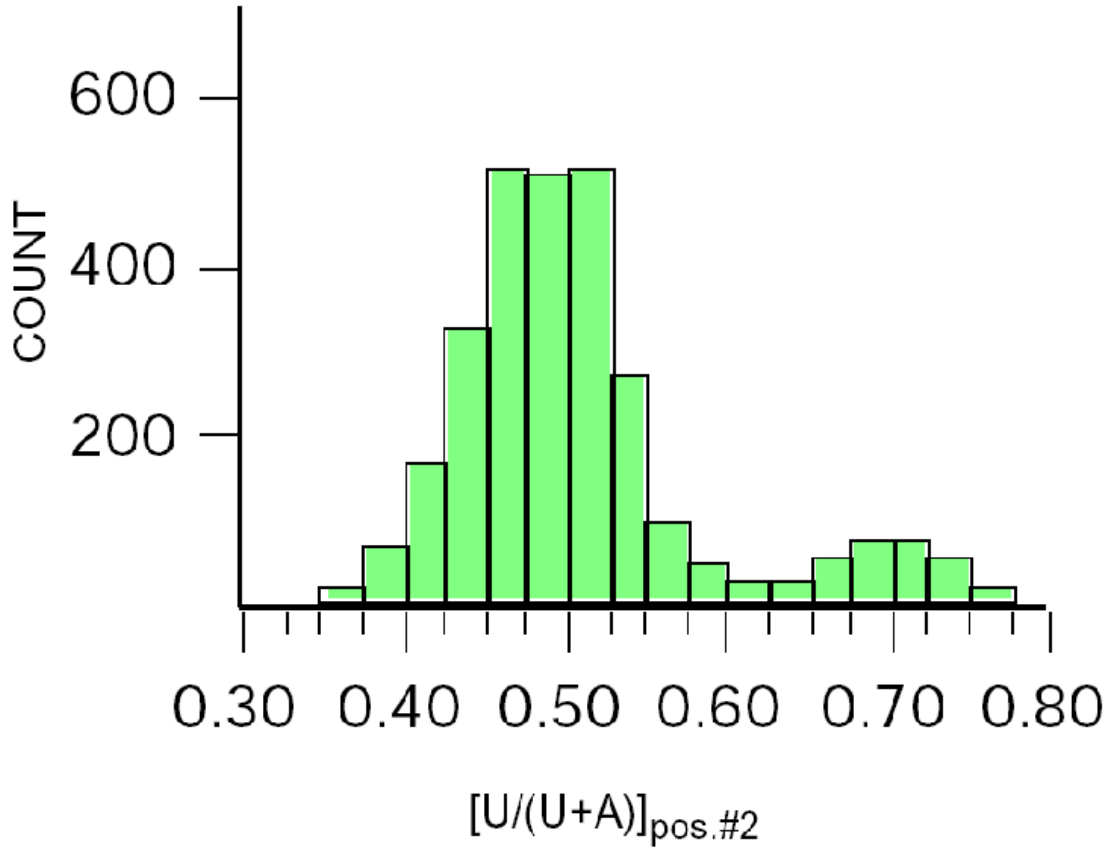


Nucleotide composition by codon position for *E. coli* genes.

Codon Usage


SECOND POSITION
T
C
A
G
THIRD POSITION
FIRST POSITION
T
C
A
G
T/C
A/G
SER
SER
TYR
CYS
STOP
STOP
TRP
LEU
PRO
PRO
HIS
GLN
SER
ASN
LYS
THR
ARG
NON POLAR
UNUSUAL
POLAR
ALA PHE
VAL MET
LEU
ILE
PRO
GLY
CYS
SER ASP TYR
THR GLU TRP
ASN HIS LYS
GLN ARG

Codon Usage



The relative content of T to (T+A) at the second position of 3180 *E. coli* genes.

Gene Prediction Software

Most gene prediction software makes use of an HMM method to simultaneously search for all these patterns as an indication of gene presence.

Hidden Markov Models (HMM)

A common way to find any pattern is to use a method now known as "Hidden Markov Models"

- It is a method based on the Markov process (named after Andrey Markov 1856-1922)
- The idea is to use a systematic model to search for a pattern from sequence data
- The pattern is hidden from you (hence you need to search for it) but the sequence data is given

Hidden Markov Models (HMM)

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

different possible "states" that the system can be in.

Hidden Markov Models (HMM)

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

different possible "states" that the system can be in.

For example, x_1 could be an intergenic region, x_2 could be the state of being in CpG island, x_3 could be the state of being in a coding region.

Hidden Markov Models (HMM)

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \begin{bmatrix} t_{11} & t_{12} & t_{13} \\ t_{21} & t_{22} & t_{23} \\ t_{31} & t_{32} & t_{33} \end{bmatrix}$$

different possible "states" that the system can be in.

For example, x_1 could be an intergenic region, x_2 could be the state of being in CpG island, x_3 could be the state of being in a coding region.

As you move linearly along a sequence there are probabilities of switching from one state to the other. That is, of moving from an 'intergenic region' state into a 'coding region' state.

Hidden Markov Models (HMM)

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

different possible "states" that the system can be in.

The trick is that these 'states' are unknown to you.

You know however that they have an influence on properties that you can observe/measure. For example you know that in a CpG island there should be more CpG's than there are in an intergenic region. You know that a coding region should have a greater level of complexity, etc.

Hidden Markov Models (HMM)

Hidden Markov models are widely used whenever some pattern needs to be found. Within sequence anlaysis common uses are for ...

- genefinding
- profile searches
- sequence alignment
- regulatory site identification

Hidden Markov Models (HMM)

We will consider a **Toy** example of an HMM given by Sean Eddy in Nature Biotechnology 22:1315 2004. The goal is to find a $5'$ splice site given just the sequence data and some properties of introns/exons. The properties are ...

- introns are AT rich
- exons are complex sequences
- splice sites are mostly G, occasionally A

Hidden Markov Models (HMM)

The properties are ...

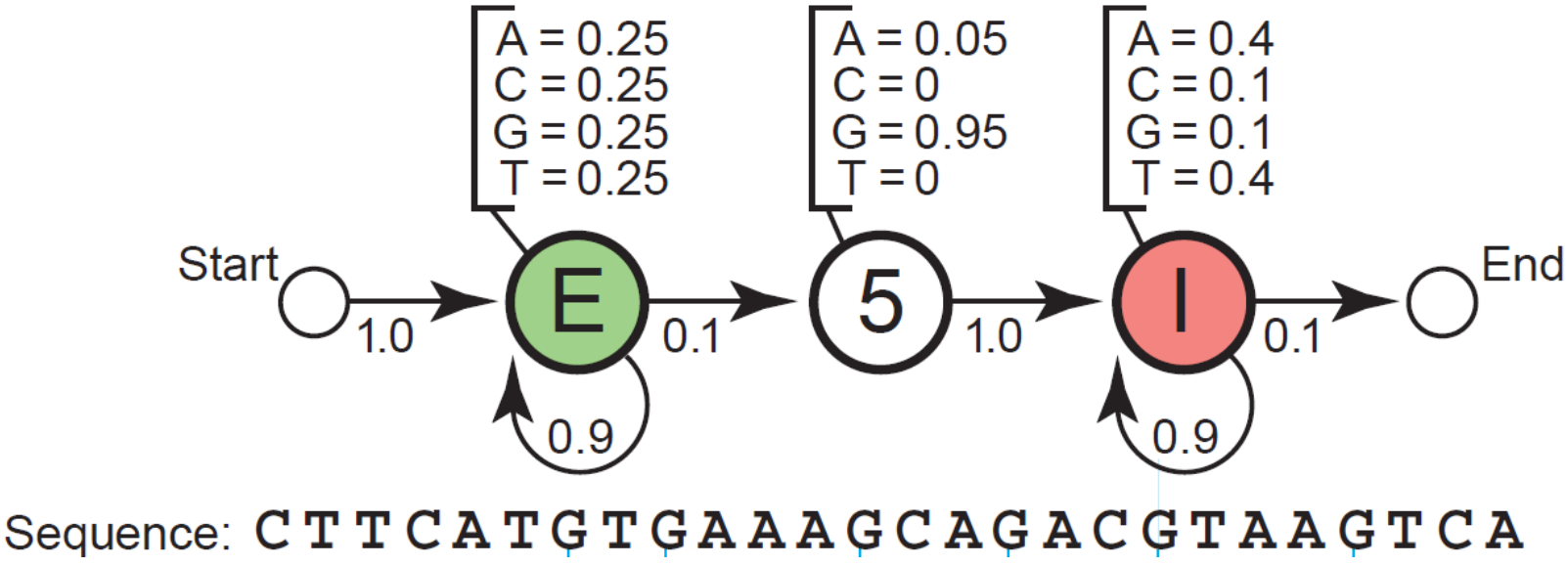
- introns are AT rich
- exons are complex sequences
- splice sites are mostly G, occasionally A

To make these more explicit ...

introns have	0.40 A	0.10 C	0.10 G	0.40 T
exons have	0.25 A	0.25 C	0.25 G	0.25 T
splice sites are	0.05 A	0.00 C	0.95 G	0.00 T

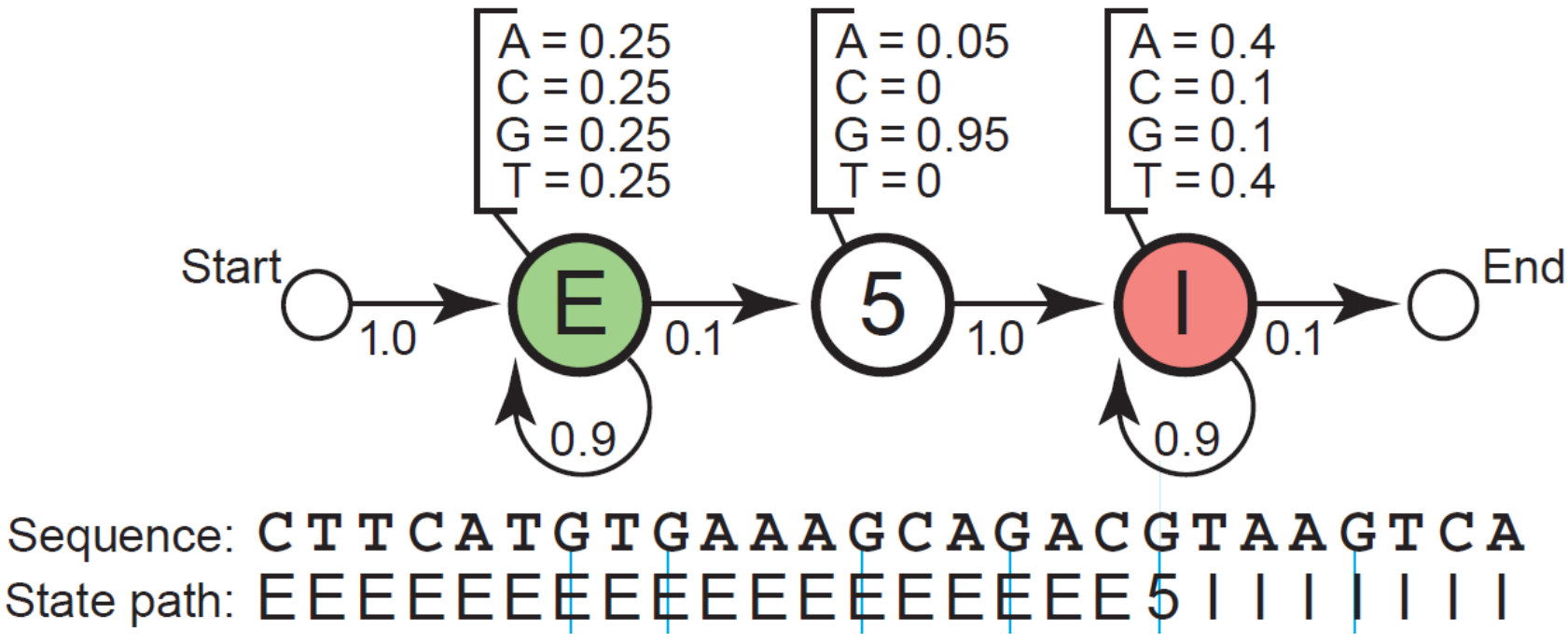
Start with a model for the pattern ...

Hidden Markov Models (HMM)

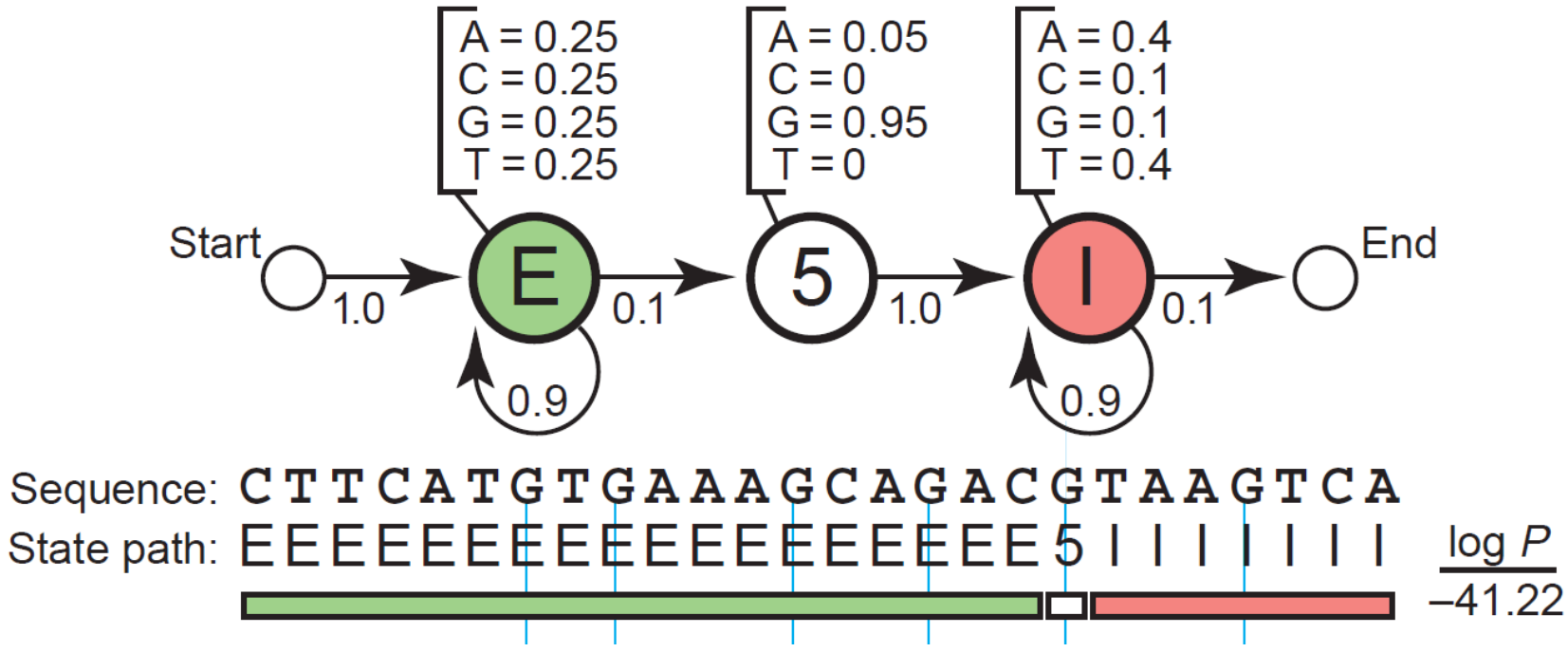


The model has both **Transition probabilities** (e.g. from state to state; from E to 5 to I) and **Emission probabilities** (chances of 'seeing' a particular nucleotide given the state).

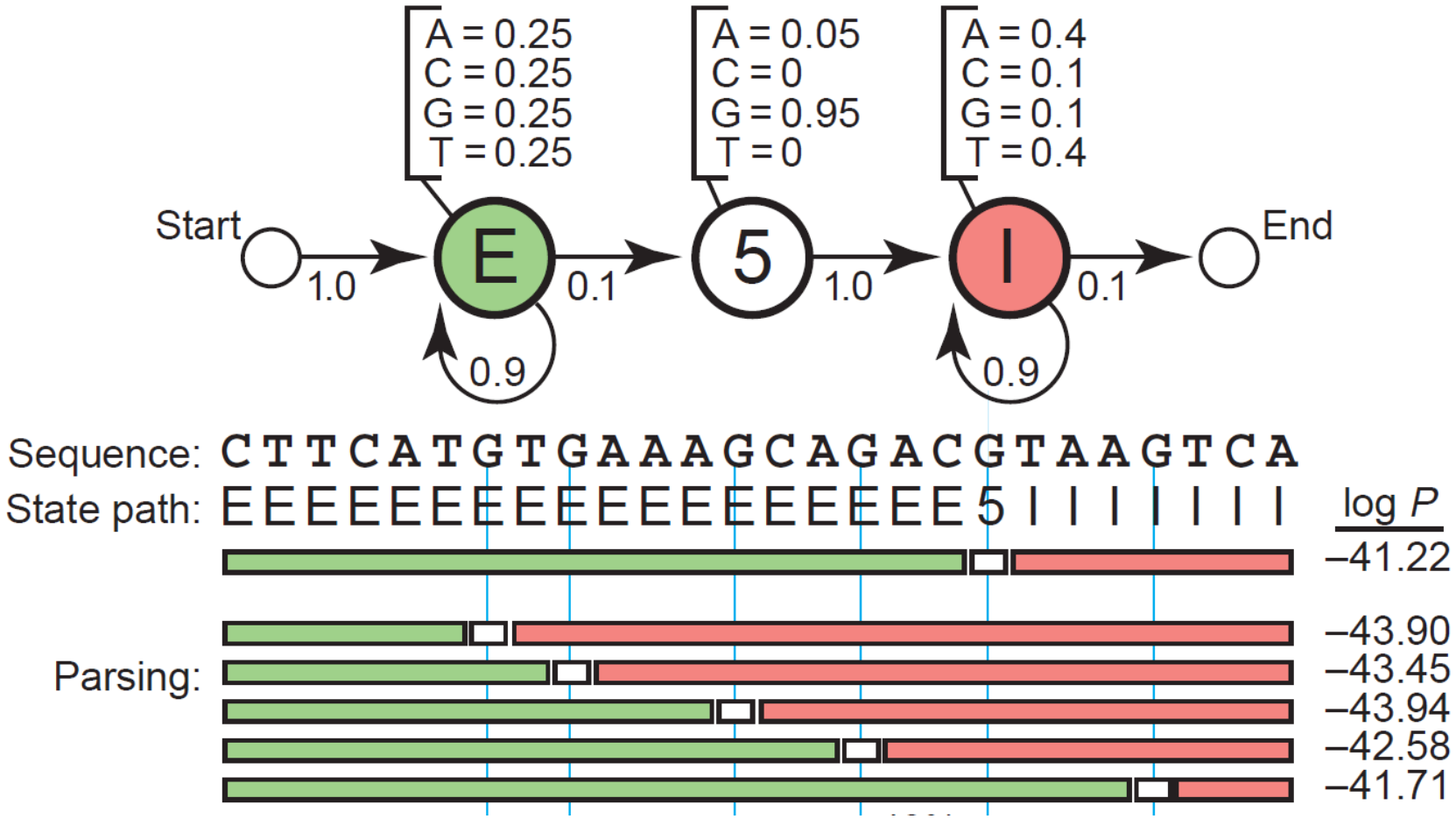
Hidden Markov Models (HMM)



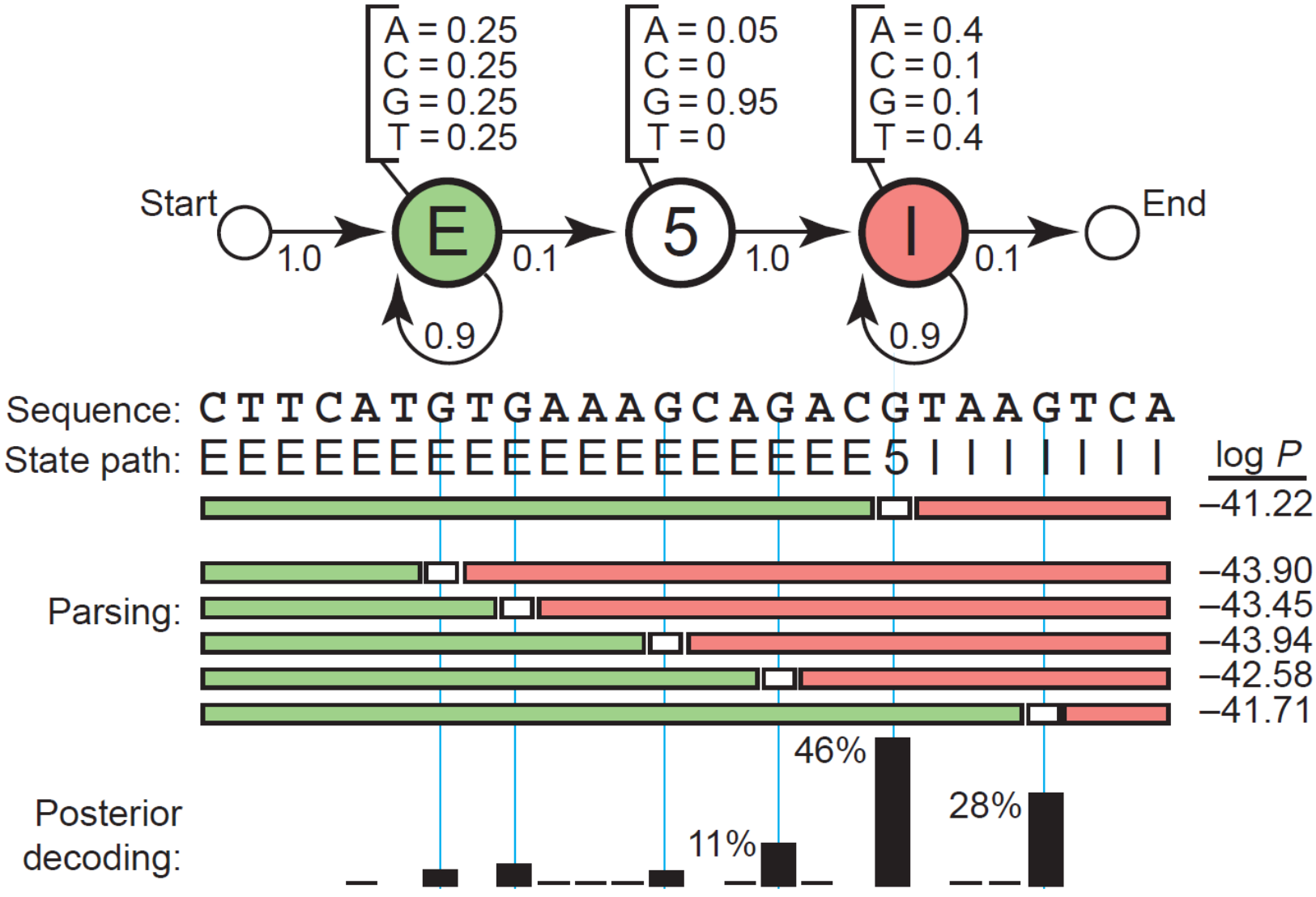
Hidden Markov Models (HMM)



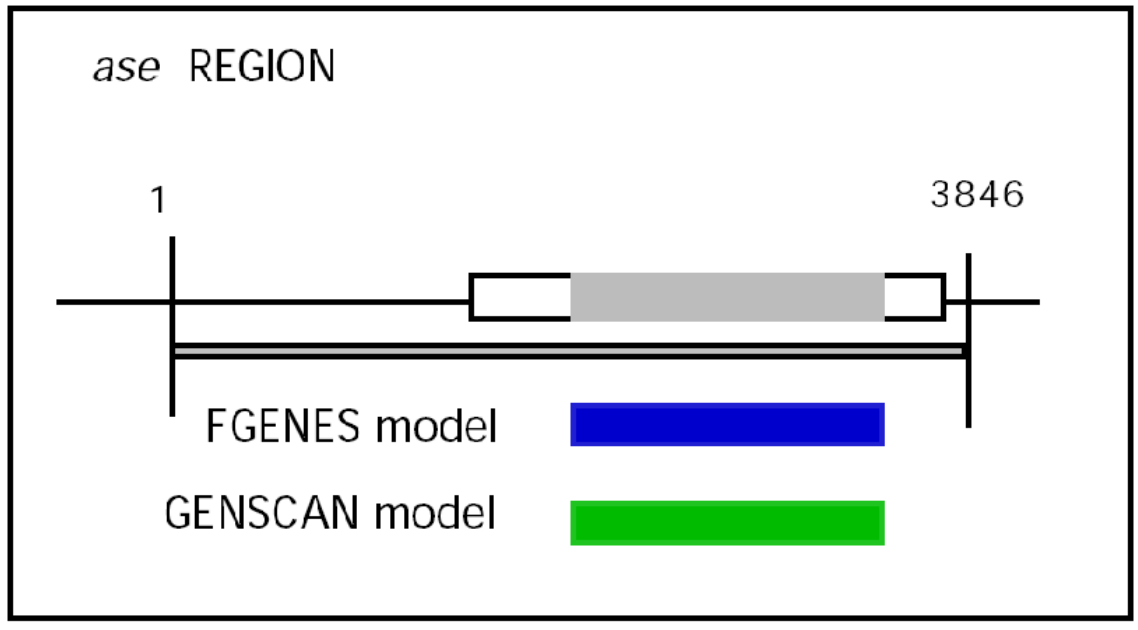
Hidden Markov Models (HMM)



Hidden Markov Models (HMM)

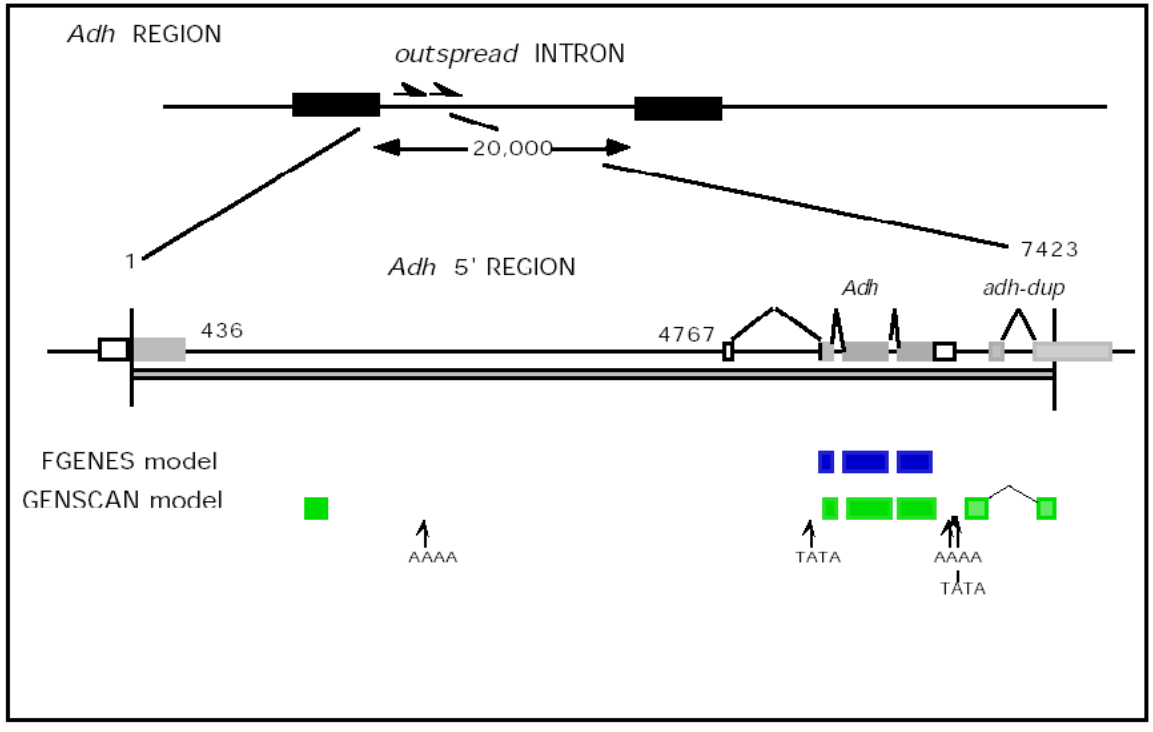


Comparison of Programs



Gene models of the *D. melanogaster* ase region.

Comparison of Programs



Gene models of the *D. melanogaster* Adh region.